J Transp Secur (2008) 1:81–106
DOI 10.1007/s12198-007-0006-4

Investigating training, transfer and viewpoint effects resulting from recurrent CBT of X-Ray image interpretation

Saskia M. Koller · Diana Hardmeier · Stefan Michel · Adrian Schwaninger

Received: 26 October 2007 / Accepted: 1 November 2007 /
Published online: 9 January 2008

Abstract X-ray screening of passenger bags is an essential task at airport security checkpoints. In this study we investigated how well airport security screeners can detect guns, knives, improvised explosive devices (IEDs) and other threat objects in X-ray images of passenger bags before and after 3 and 6 months of recurrent (about 20 min per week) computer-based training (CBT). Two experiments conducted at different airports gave very similar results. Training with X-ray Tutor (XRT), an individually adaptive CBT, resulted in large performance increases, especially for detecting IEDs. While performance for detecting IEDs was initially substantially lower than for guns, IEDs could be detected as well as guns after several months of training. A large transfer effect was observed as well: Training with XRT helped screeners recognize new threat objects that were similar in shape as the trained objects. Threat recognition was dependent on the rotation of the objects. If depicted from an unusual viewpoint, prohibited items were more difficult to recognize. The results were compared to two conventional (not adaptive) CBT systems. For one system no training and transfer effects were observed whereas small training and transfer effects were found for the other conventional CBT system.

Keywords Airport security · Human–computer interaction · Human factors · Object recognition · Perceptual learning · X-ray screening

S. M. Koller · D. Hardmeier · S. Michel · A. Schwaninger
Department of Psychology, University of Zurich, Binzmühlestrasse 14/22, 8050 Zurich, Switzerland

S. Michel · A. Schwaninger (✉)
Department Bülthoff, Max Planck Institute for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany
e-mail: a.schwaninger@psychologie.uzh.ch

Introduction

The importance of aviation security has increased dramatically in the last years. As a consequence of the new threat situation, large investments were made into modern security technology. State of the art X-ray screening equipment offers good image quality, high resolution and many image enhancement functions. However, the decision whether an X-ray image of a passenger bag contains a prohibited item or not, is still being taken by a human operator, i.e. an airport security screener. Object shapes that are not similar to ones stored in visual memory are difficult to recognize (e.g., Graf et al. 2002; Schwaninger 2004, 2005). Schwaninger et al. (2005) have shown that X-ray screener performance depends on knowledge-based and image-based factors. A prerequisite for good X-ray detection performance is knowledge about which objects are prohibited and what they look like in X-ray images. Such knowledge is acquired by computer-based, class-room and on the job training (knowledge-based factors). Image-based factors refer to image difficulty resulting from viewpoint variation of threat objects, superposition of threat objects by other objects in a bag, and bag complexity depending on the number and type of other objects in the bag. The ability to cope with image-based factors is related to individual visual-cognitive abilities rather than a mere result of training (Hardmeier et al. 2006).

Computer-based training is expected to be a very important determinant of X-ray image interpretation competency, because many threat objects are not known from everyday experience and because objects look quite different in X-ray images than in reality. This is illustrated in Fig. 1 with two examples.

Schwaninger and Hofer (2004) and Schwaninger et al. (2007) could show that detection of improvised explosive devices (IEDs) in hold baggage screening (HBS) can be significantly improved if people are trained with an individually adaptive training system such as X-ray Tutor (XRT). Schwaninger et al. (2005) compared detection performance of novices with the one of aviation security screeners. A rather poor recognition of unfamiliar object shapes (e.g., self-defense gas spray, electric shock device etc.) in x-ray images was found for novices. For experienced aviation security personnel, a much higher recognition performance was observed. McCarley et al. (2004) reported a better performance after training for the detection of knives in X-ray images for novices.

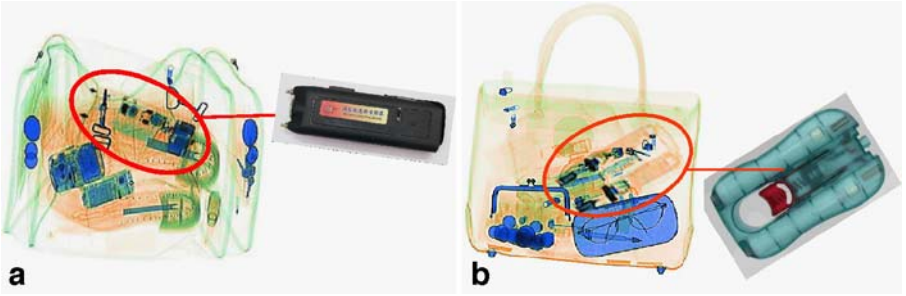

Fig. 1 Different types of prohibited items in X-ray images of passenger bags. **a** Electric shock device, **b** self defense gas spray "Guardian Angel"

When one takes into account the myriad of views that can be produced by a single object, the question arises how the human brain stores and recognizes objects even if they are presented in unusual views. In the object recognition literature, two types of theories can be distinguished: structural description theories and view-based theories. The former assume that objects are stored in visual memory by their component parts and their spatial relationship. An object-centered description of this nature was described by Marr and Nishihara (1978), who proposed that objects are hierarchically decomposed into their parts and spatial relations relative to object-centered coordinates in order to access an object-centered 3D model in visual memory. In Biederman (1987) recognition by components (RBC) theory, non-accidental properties like vertices, parallel vs. non-parallel lines, straight vs. curved lines etc. (see Lowe, 1985, 1987) are extracted from a line drawing representation of objects to define basic geometrical primitives (geometrical ions, "geons") that are relatively orientation-invariant. A geon structural description (GSD) in memory is activated by extracting geons from the visual input and match geon properties and their spatial relationship with the GSD (Hummel and Biederman 1992).

For view-based theories, different approaches have been proposed. Examples are recognition by alignment to a 3D representation (Lowe 1987), recognition by linear combination of 2D views (Ullman 1998), recognition by view interpolation (e.g., using RBF networks) proposed by Poggio and Edelman (1990) and storing of multiple views for each object plus performing transformations (Tarr and Pinker 1989). What view-based theories have in common is the assumption that objects are not stored in memory as rotation invariant structural descriptions but instead in a format which is viewer-centered. A more detailed discussion of structural description theories vs view-based theories and more recent hybrid theories is beyond the scope of this paper (for reviews see for example Graf et al. 2002; Hayward 2003; Kosslyn 1994; Peissi and Tarr 2007; Schwaninger 2005; Tarr and Bülthoff 1998). However, it should be pointed out that empirical results seem to be correlated with the required level of recognition (Bülthoff et al. 1995, p. 5, 13; Tarr 1995): if the object has to be recognized at 'entry level', behavioral measures are less affected by changes in perspective. However, in the case of subordinate recognition in which fine discriminations are typically required, both response times and accuracy are more sensitive to the specific viewpoint used. Furthermore, differences in the task a subject has to perform (Lawson 1999) and the specific paradigm that is used (Verfaillie 1992) can influence which level of representation is tapped (see also Logothetis and Sheinberg 1996).

The first aim of this study is to investigate how well airport security screeners can detect guns, knives, IEDs and other prohibited items in x-ray images of passenger bags. The second aim is to examine whether screener detection performance can be increased by conducting recurrent CBT. To this end, screeners conducted weekly recurrent CBT (about 20min per week). Detection performance was tested with the X-ray Competency Assessment Test (X-ray CAT) by Koller and Schwaninger (2006). This test measures how well people detect threat items in X-ray images of passenger bags. It was conducted at the beginning and then after 3 and 6months of training. In addition to training effects, the X-ray CAT allows measuring transfer effects, i.e. to what extent visual knowledge that was gained through CBT can be transferred to other threat items (see below). In the X-ry CAT all prohibited items are

depicted from a canonical (easy recognizable) perspective (Palmer et al. 1981) and unusual perspective which allows investigating viewpoint effects. The study was conducted at two mid-size European airports. In airport 1 (experiment 1) one group of screeners used adaptive CBT (XRT) whereas the other group of screeners (control group) used a conventional (not adaptive) CBT. In airport 2 (experiment 2) the same experimental design was used except for the fact that the control group used another conventional CBT system. This allows investigating whether a training effect is dependent on the type of the CBT system used.

Experiment 1

Method

Participants

A total of 209 airport security screeners of a mid-size European airport participated in experiment 1 and conducted the X-ray CAT 1.0.0 three times with an interval of 3 months between the measurements. The adaptive CBT group (XRT group) consisted of 97 screeners who conducted weekly recurrent CBT using X-ray Tutor (XRT) CBS 2.0 Standard Edition between all three test measurements. The control group consisted of 112 screeners who used a conventional (not adaptive) CBT. According to the security organization and their appropriate authority, airport security screeners of both groups conducted about 20 min CBT per week. Analysis of XRT training use showed that on average, each screener trained 20.26 min (SD = 3.65 min) per week.

Materials and procedure

X-Ray competency assessment test (X-ray CAT)

The X-ray CAT consists of 256 trials based on 128 different color X-ray images of passenger bags. Each of the bag images is used once containing a prohibited item (threat image) and once without any threat object (non-threat image). Figure 2 displays examples of the stimuli. Note that in the test, the images are displayed in color.

Fig. 2 Example images from the X-ray CAT. Left: harmless bag (non-threat image), *right* same bag with a prohibited item at the *top right corner* (threat image). The prohibited item (gun) is shown also separately *at the bottom right*

Prohibited objects can be assigned to four categories as defined in Doc 30 of the European Civil Aviation Conference (ECAC): guns, IEDs, knives and other prohibited items (e.g., self-defense gas spray, chemicals, grenades, etc.). The threat objects have been selected and prepared in collaboration with experts of Zurich State Police, Airport Division to be representative and realistic. For each threat category 16 exemplars are used (eight pairs). Each pair consists of two prohibited items that are similar in shape (see Fig. 3). These were distributed randomly into two sets, sets A and B. Prohibited items of set A (non threat bag images) are contained in the XRT CBS 2.x SE training whereas the items of set B are not. This allows testing for transfer effects.

Every item is depicted from two different viewpoints. The easy viewpoint refers to the canonical (i.e. easy recognizable) perspective (Palmer et al. 1981). The difficult viewpoint shows the threat item with an 85° horizontal rotation or an 85° vertical rotation relative to the canonical view (see Fig. 3 for examples). In each threat category, half of the prohibited items of the difficult viewpoint are rotated vertically, the other half horizontally. Sets A and B are equalized concerning the rotations of the prohibited objects.

Every threat item is combined with a bag in a manner that the degree of superposition by other objects is similar for both viewpoints. This was achieved using a function that calculates the difference between the pixel intensity values of the bag image with the threat object minus the bag image without the threat object using the following formula:

$$SP = \frac{\sqrt{\sum [I_{SN}(x, y) - I_N(x, y)]^2}}{ObjectSize}$$

SP = Superposition; I_{SN} = grayscale intensity of the SN (signal plus noise) image (contains a prohibited item); I_N = grayscale intensity of the N (noise) image (contains no prohibited item); Object Size: Number of pixels of the prohibited item where *R*, *G* and *B* are <253

Using this equation (division by object size), the superposition value is independent of the size of the prohibited item. This value can be kept relatively constant for the two views of a threat object, independent of the degree of clutter in a bag, when combining the bag image and the prohibited item. The bag images were visually inspected by aviation security experts to ensure they do not contain any

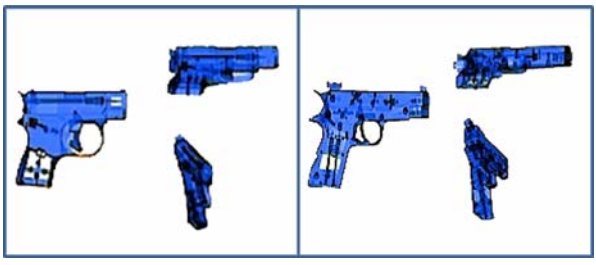

Fig. 3 Example of two X-ray images of similar looking threat objects used in the test. *Left* A gun of set A. *Right* Corresponding gun of set B

other prohibited items. Harmless bags were assigned to the different categories and viewpoints of the threat objects in a way that their difficulty was balanced across all categories.[1] The false alarm rate (the rate at which screeners wrongly judged a harmless bag as containing a threat item) for each bag image served as measure of difficulty based on a pilot study with 192 screeners of another airport.

The X-ray CAT takes about 30–40 min to complete. Each image is shown for a maximum of 10 s on the screen. Screeners have to judge whether the bag is OK (contains no prohibited item) or Not OK (contains a prohibited item). Additionally, screeners have to indicate the perceived difficulty of each image on a 100-point scale (difficulty rating).[2] The X-ray CAT is built into the XRT training system (see below). The interface of the X-ray CAT is the same as in XRT except there is no feedback and screeners do not have to click on the image to identify the threat object.

X-Ray tutor (XRT) training system

X-Ray Tutor (XRT) is an individually adaptive training system for aviation security screeners. It contains a large image library with hundreds of different threat objects depicted in up to 72 views, more than 6,000 bag images and many millions of possible threat object to bag combinations (see Schwaninger 2004 for details). The individually adaptive training algorithm of XRT starts with showing threat objects depicted from easy viewpoints with little superposition by other objects and in bags of low complexity. Based on each individual screeners' learning progress, threat objects are shown in more difficult views, more complex bags and with more superposition. These parameters are adapted automatically by a scientifically validated algorithm for each screener and threat object while taking into account automatic image processing algorithms as explained in Schwaninger et al. (2007). XRT first presents screeners prohibited objects in easy (canonical) views. The individually adaptive training algorithm determines for each screener which views are difficult to recognize and adapts the training so that the trainee becomes able to detect threat items reliably even if prohibited objects are substantially rotated away from the easiest view. During the next difficulty levels, first superposition and then bag complexity is increased so that the trainee becomes able to detect threat items reliably even if they are superimposed by other objects or if the complexity of a bag is very high (for more information on XRT see Schwaninger 2003, 2004, and 2005a).

During a training session each image is displayed for 15 s on the screen. Within this time screeners can use image enhancement functions which are also available when working with the X-ray machine (e.g. grayscale, negative image, edge enhancement, etc.). If the image contains a prohibited item, screeners have to click on it and then click on the Not OK button. If the bag is harmless; they have to click on the OK button. After providing a confidence rating using a slider control, feedback is shown to inform the trainee whether the image has been judged correctly

[1] The eight categories of test images (four threat categories in two viewpoints each) are similar in terms of the difficulty of the harmless bags. This means, a difference of detection performance between categories or viewpoints can not be due to differences in the difficulty of the bag images.

[2] The difficulty ratings were not analyzed in this study.

or not (see Fig. 4). If the bag contains a threat item, it is highlighted by flickering and the trainee has the possibility to display information about the threat item (see bottom left of Fig. 4). By clicking on the continue button the next image is shown. As a default setting, one training sessions takes 20min. During this time screeners see between 150 and 300 images.

Procedure

As explained above, two groups of screeners participated in experiment 1. The XRT training group conducted weekly recurrent CBT using XRT CBS 2.0 Standard Edition. The control group used a conventional (not adaptive) CBT. In order to avoid potential negative consequences, we decided not to mention the exact CBT product in this article. However, it can be mentioned that this CBT is also widely used at many airports worldwide. It has a much smaller threat image library than XRT, threat objects are not displayed in many different views, threat objects are not matched with different bags on the fly, and there is no individually adaptive training algorithm.

The XRT training group and the control group took the X-ray CAT before, after three, and after six months of weekly CBT. This allows testing the effectiveness of both CBT systems for increasing X-ray image interpretation competency of airport security screeners. As explained above, half of the prohibited items in the X-ray CAT are also contained in the XRT training system (although presented in different

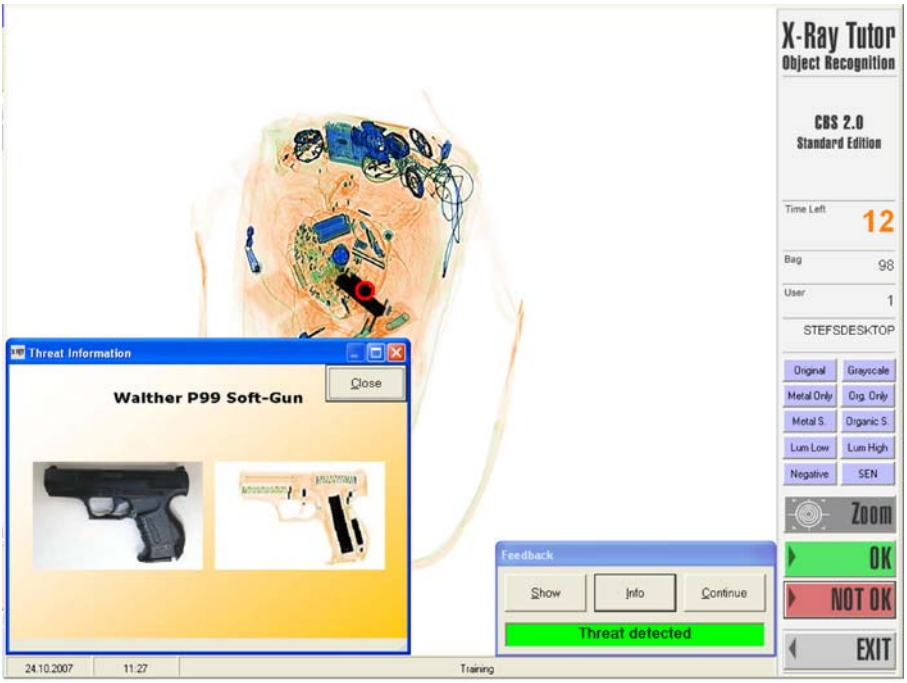

Fig. 4 Screenshot of the XRT CBS 2.0 training system during training. *At the bottom right* a feedback is provided after each response. If a bag contains a prohibited item, an information window can be displayed (see *bottom left of the screen*)

bags). The other half of the prohibited items of the X-ray CAT are not part of the XRT training library. This allows testing for transfer effects, i.e. testing whether training with the detection of certain prohibited items helps increasing the detection of other prohibited items. Finally, as specified above in the section on the X-ray CAT, all prohibited items are depicted in easy and difficult view which allows testing effects of viewpoint on screener detection performance.

Results and discussion

Detection performance was calculated using the signal detection measure d' (Green and Swets 1966), which takes into account the hit rate (correctly judged threat images as being Not OK) and the false alarm rate (wrongly judged harmless bags as being Not OK). D' is calculated using the following formula:

$$\mathrm{d}' = \mathrm{z(H)} - \mathrm{z(FA)}$$

Whereas H is the hit rate, FA the false alarm rate and z refers to the z transformation. Performance values are not reported due to security reasons. However, effect sizes are reported for all relevant analyses and interpreted based on Cohen (1988), see Table 1. For t tests, d between 0.20 and 0.49 represents small effect size; d between 0.50 and 0.79 represents medium effect size; $d \geq 0.80$ represents large effect size. For analysis of variance (ANOVA) statistics, η^2 between 0.01 and 0.05 represents small effect size; η^2 between 0.06 and 0.13 represents medium effect size; $\eta^2 \geq 0.14$ represents large effect size.

Figure 5 shows the detection performance of the first, second and third measurement for both screener groups. As can be seen in the Figure, there was a large improvement as a result of training in the XRT training group while there was no improvement in the control group. These results were confirmed by an ANOVA for repeated measures using d' scores with the within-participant factor measurement (first, second and third) and the between-participants factor group (XRT training group and control group). There were large main effects of measurement, $\eta^2 = 0.28$, $F(2, 414) = 81.04$, $p < 0.001$, and group, $\eta^2 = 0.19$, $F(1, 207) = 47.62$, $p < 0.001$. There was also a large interaction of measurement and group, $\eta^2 = 0.25$, $F(2, 414) = 68.67$, $p < 0.001$, which is consistent with Fig. 5 showing large performance increases as a result of training only for the XRT training group but not for the control group.

Separate pairwise t tests of detection performance d' revealed no significant difference at the baseline measurement between the two groups $t(177) = -0.91$, $p = 0.363$, $d = 0.13$, but already a significant difference in the second measurement, i.e. after three months of training, $t(207) = 7.52$, $p < .001$, $d = 1.04$. Additional paired-samples t tests revealed significant differences for the XRT training group between all three test measurements but no significant differences for the control group (see Table 2).

Table 1 Classification of effect sizes based on Cohen (1988)

Effect size	d	H^2
Small	0.20–0.49	0.01–0.05
Medium	0.50–0.79	0.06–0.13
Large	≥0.8	≥0.14

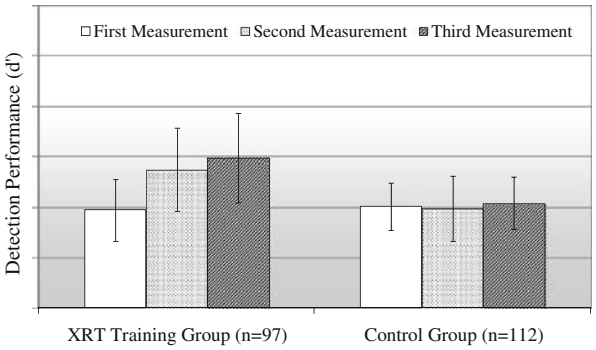

Fig. 5 Detection performance with standard deviations for the XRT training group (*left*) vs the control group (*right*) comparing first, second and third measurement

Figure 6 shows the detection performance of both screener groups broken up by prohibited item category and the three test measurements. A repeated-measures ANOVA with the within-participant factors measurement (first, second and third) and threat category (guns, IEDs, knives and other), and the between-participants factor group (XRT training vs control) revealed the significant main effects and significant interactions given in Table 3, a. In addition to the effects that were already found in the previous ANOVA, also the factor threat (or prohibited item) category was significant. As can be seen in Fig. 6, guns were detected best, followed by knives, other prohibited items and IEDs at the first test measurement. There was a highly significant interaction between threat category and measurement. As can be seen in Fig. 6, detection of IEDs was initially much lower than gun detection. After 6 months of training, screeners of the XRT training group could detect IEDs even slightly better than guns. This result implies that IED detection is not difficult per se but rather a matter of the right training. Note that in this study all IEDs contained a detonator, wires, explosive, a triggering device and a power source. Thus our conclusions are only applicable to the detection of such multi-component IEDs. Large performance increases were also found for other prohibited items in this group, while for knives, only a small improvement as a result of training was found. Note that after 6months of training, detection performance of knives is lower than the one for any other threat category in the XRT training group, although at baseline measurement it was higher than the detection performance for IEDs or other threat objects. The interaction between threat category, group and measurement is also worth mentioning. As can be seen in Fig. 6 this results from the fact that there was no training effect for the control group. Their detection performance remains at about the same level for each threat category even after 6months of training with the conventional (not adaptive) CBT system.

Separate pairwise *t* tests were conducted to compare detection performance at the first and the second measurement for both groups and each threat category separately

Table 2 Results of the *t* tests comparing the detection performance of first (*t*1), second (*t*2) and third (*t*3) measurement

	t(96)	*t*(111)	*p*	*D*
XRT training group (*t*1–*t*2)	−9.80		<0.001	1.12
XRT training group (*t*2–*t*3)	−3.95		<0.001	0.28
Control group (*t*1–*t*2)		0.54	=0.59	0.05
Control group (*t*2–*t*3)		−1.89	=0.06	0.17

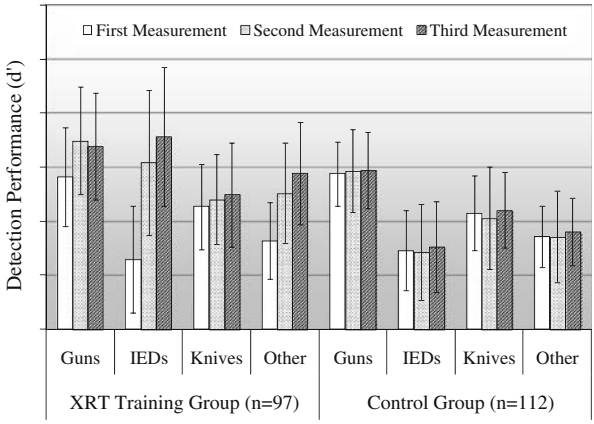

Fig. 6 Detection performance with standard deviations for the XRT training group vs the control group broken up by prohibited item category and test measurement

(Table 4). The XRT training group showed a significant increase of the detection performance at the second measurement for the categories guns, IEDs and other threat objects. For knives, a significant difference could be found only in the third measurement. The comparison of the effect size *d* between the *t* tests of the four threat categories confirms the earlier mentioned conclusion that the training effect was particularly big for IEDs and rather small for knives. Detection performance of the control group did not differ significantly between the measurements, confirming that the conventional CBT did not result in an increase of threat detection performance.

The results of the analyses considering the two prohibited item sets of the X-ray CAT, set A and set B, are shown in Figs. 7 and 8. As explained above, set A are X-ray CAT images which contain prohibited items which are part of the XRT image library. Set B are X-ray CAT images which contain prohibited items that are not part of the XRT image library. By comparing training effects for sets A and B transfer effects can be investigated, i.e. whether training with XRT does not only improve detection of prohibited items that are part of the XRT image library (set A) but also the detection of other prohibited items that are visually similar (set B). Figure 7 shows the detection performance for both screener groups broken up by test set for all three measurements. It shows a clear increase in detection performance for the XRT training group, especially at the second measurement, after the first 3 months of training. For the control group, as in the previous analysis, no training effect is evident. The results of the repeated measures ANOVA with the within-participant factors measurement (first, second and third) and set (A vs B) and the between-participant factor group (XRT training group vs. control group) can be seen in Table 3, b. There was a significant effect of set in this analysis, which would imply a different detection performance for set A vs set B. However, the effect is very small, as the effect size of $\eta^2 = 0.2$ clearly shows, which makes the difference quasi negligible. This is also supported by the small effect size for the interaction between set and measurement, $\eta^2 = 0.4$. Pairwise *t* tests showed a significant increase in detection performance at the second measurement for both sets for the XRT training group, set A, $t(96) = -10.27$, $p < .001$, $d = 1.19$, set B, $t(96) = -7.68$, $p < .001$, $d = 0.92$. These results indicate a large transfer effect, i.e. visual knowledge regarding the visual appearance of the prohibited objects of the XRT image library helped screeners to detect similar looking, but untrained objects in the X-ray CAT

Table 3 Results of the ANOVAs in experiment 1

	Factor	*df*	*F*	η^2	*p* Value
a	Measurement (*M*)	2, 414	83.96	0.29	<0.001
	Threat category (*T*)	3, 621	240.03	0.54	<0.001
	Group (*G*)	1, 207	56.20	0.21	<0.001
	*M*x*G*	2, 414	70.49	0.25	<0.001
	*T*x*G*	3, 621	45.05	0.18	<0.001
	*M*x*T*	6, 1242	43.20	0.17	<0.001
	*M*x*T*x*G*	6, 1242	40.65	0.16	<0.001
b	Measurement (*M*)	2, 414	80.55	0.28	<0.001
	Set (*S*)	1, 207	4.18	0.02	<0.05
	Group (*G*)	1, 207	49.40	0.19	<0.001
	*M*x*G*	2, 414	67.99	0.25	<0.001
	*M*x*S*	2, 414	8.80	0.04	<0.001
	*S*x*G*	1, 207	51.32	0.20	<0.001
	*M*x*S*x*G*	2, 414	11.54	0.05	<0.001
c	Measurement (*M*)	2, 414	87.69	0.30	<0.001
	Set (*S*)	1, 207	2.37	0.01	=0.13
	Threat category (*T*)	3, 621	236.79	0.53	<0.001
	Group (*G*)	1, 207	63.57	0.24	<0.001
	*M*x*G*	2,414	71.16	0.26	<0.001
	*M*x*T*	6, 1242	44.35	0.18	<0.001
	*M*x*S*	2, 414	10.93	0.05	<0.001
	*S*x*G*	1, 207	52.25	0.20	<0.001
	*S*x*T*	3, 621	74.00	0.26	<0.001
	*T*x*G*	3, 621	47.39	0.19	<0.001
	*M*x*T*x*G*	6, 1242	41.04	0.17	<0.001
	*M*xSx*G*	2, 414	10.74	0.05	<0.001
	*M*x*S*x*T*	6, 1242	3.84	0.02	<0.01
	*S*x*T*x*G*	3, 621	4.78	0.02	<0.01
	*M*x*S*x*T*x*G*	6, 1242	2.99	0.01	<0.01
d	Measurement (*M*)	2,414	84.10	0.29	<0.001
	View (*V*)	1, 207	1768.63	0.90	<0.001
	Threat Category (*T*)	3, 621	258.62	0.56	<0.001
	Group (*G*)	1, 207	61.91	0.23	<0.001
	*M*x*G*	2, 414	65.80	0.24	<0.001
	*M*x*T*	6, 1242	41.33	0.17	<0.001
	*M*x*V*	2, 414	2.05	0.01	=0.13
	*V*x*G*	1, 207	3.27	0.02	=0.07
	*V*x*T*	3, 621	425.64	0.67	<0.001
	*T*x*G*	3, 621	40.86	0.17	<0.001
	*M*x*T*x*G*	6, 1242	40.25	0.16	<0.001
	*M*x*V*x*G*	2, 414	2.23	0.01	<0.05
	*M*x*V*x*T*	6, 1242	6.58	0.03	<0.001
	*V*x*T*x*G*	3, 621	3.08	0.02	<0.05
	*M*x*V*x*T*x*G*	6, 1242	2.68	0.01	<0.05

(set B). Consistent with previous analyses, there was no training effect for the control group, neither for set A, $t(111) = 0.76$, $p = 0.45$, $d = 0.08$, nor for set B, $t(111) = -0.28$, $p = .78$, $d = 0.03$. Pairwise *t* tests comparing both sets within one group at the first measurement revealed a significant difference of the two sets only for the control group $t(111) = -2.82$, $p < .01$, $d = 0.17$ but not for the XRT training group, $t(96) = -0.42$, $p = .68$, $d = 0.03$. However, note that an effect size of $d = 0.17$ is very small which supports the assumption that the two sets are in fact very similar in their difficulty level.

Table 4 Results of the *t* tests comparing the detection performance of the four categories between the first (*t*1), second (*t*2) and third (*t*3) measurement

	t(96)	*t*(111)	*df*	*p* Value	*d*
XRT training group					
Guns *t*1–*t*2	−5.96		96	<0.001	0.70
IEDs *t*1–*t*2	−13.03		96	<0.001	1.53
Knives *t*1–*t*2	−1.51		96	=0.13	0.17
Other *t*1–*t*2	−8.47		96	<0.001	1.07
Guns *t*1–*t*3	−4.69		96	<0.001	0.60
IEDs *t*1–*t*3	−15.88		96	<0.001	2.00
Knives *t*1–*t*3	−2.27		96	<0.05	0.26
Other *t*1–*t*3	−12.56		96	<0.001	1.51
Control group					
Guns *t*1–*t*2		−0.40	111	=0.69	0.05
IEDs *t*1–*t*2		0.03	111	=0.98	0.00
Knives *t*1–*t*2		0.83	111	=0.41	0.09
Other *t*1–*t*2		−0.17	111	=0.87	0.02
Guns *t*1–*t*3		−0.92	111	=0.36	0.10
IEDs *t*1–*t*3		−1.05	111	=0.30	0.08
Knives *t*1–*t*3		−0.73	111	=0.47	0.08
Other *t*1–*t*3		−1.39	111	=0.17	0.15

Figure 8 includes also the threat category in the analysis. The increase in detection performance for the XRT training group can also be seen in the different threat categories. Pairwise *t* tests between the first and second measurement confirmed a significant ($p < .001$, all $d > 0.62$) increase in detection performance for the XRT training group for all threat categories per set except for knives (set A: $p = .12$, $d = 0.19$, set B; $p = .32$, $d = 0.12$). In Fig. 8, detection performance in set A for guns shows a decrease between the second and third measurement. However, this difference was not significant ($p = .13$, $d = 0.17$). For the control group, detection performance between the first and third measurement was compared in order to maximize the chances for finding a significant training effect. Even here, for all categories in each set, the detection between the first and third measurement did not differ significantly (all $p > .12$, $d < 0.18$).

The extended ANOVA with the additional within-participant factor threat category revealed the main effects and interactions as specified in Table 3, c. The main effect of set was not significant but there were significant interactions with set (see Table 3, c). However, as can be seen in Fig. 8, these interactions are rather small, which implies large transfer effects.

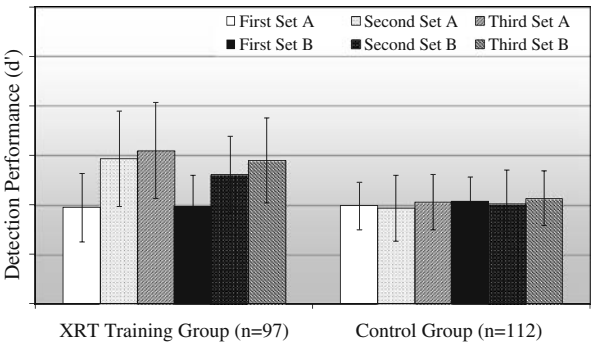

Fig. 7 Detection performance with standard deviations for the XRT training group vs the control group comparing first, second and third measurement for set A and set B separately

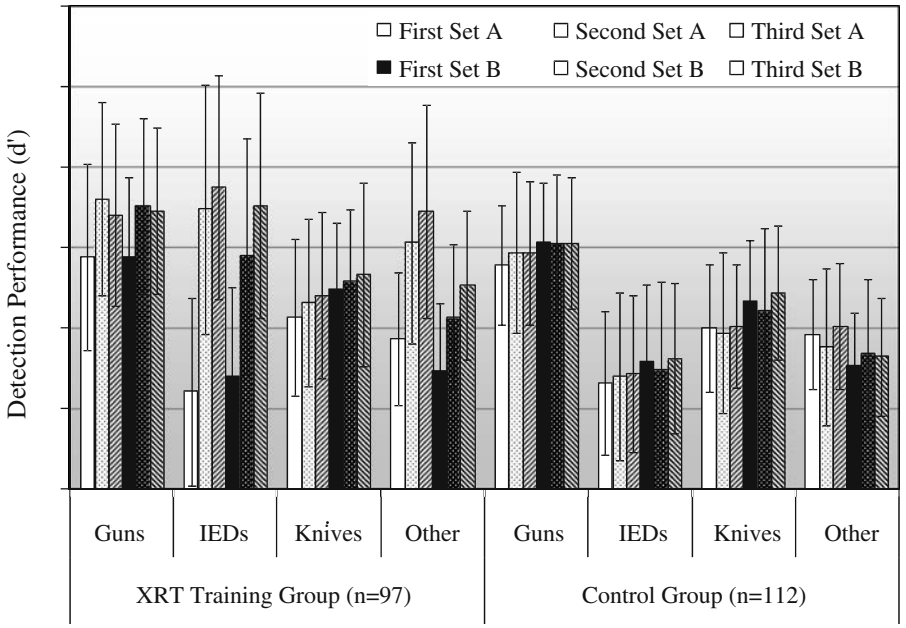

Fig. 8 Detection performance with standard deviations for the XRT training group vs the control group comparing first, second and third measurement for sets A and B and each threat category separately

Figure 9 shows the results of the viewpoint analysis. An ANOVA was conducted on d′ scores with the within-participant factors measurement, threat category and viewpoint and the between-participants factor group. It showed significant main

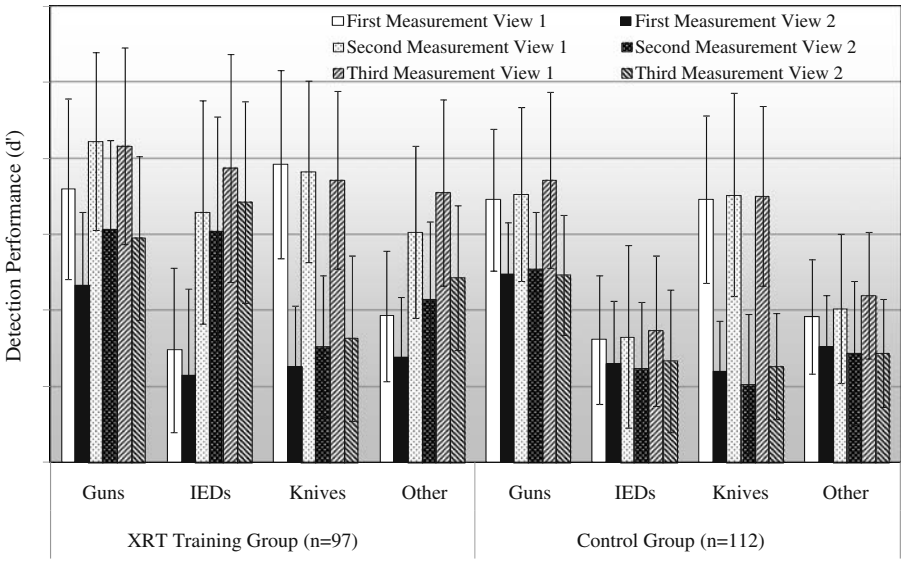

Fig. 9 Detection performance with standard deviations for the XRT training group vs the control group comparing first, second and third measurement for both views and each threat category separately

effects of measurement, category, viewpoint and group. For details and interactions see Table 3, d. The large main effect of viewpoint indicates a higher detection performance for objects in easy (canonical) viewpoint compared to objects presented in a difficult (rotated) view (cf. Fig. 9). However, no significant interaction between viewpoint and training could be found. This would suggest that the viewpoint effect is unaffected by the training and could not be decreased. Pairwise t tests showed a significant increase in detection performance at the second measurement for both views in all categories for the XRT training group with the exception of knives in the easy view ($p = .53$, $d = 0.07$). All other comparisons were significant $p < .05$, $d > 0.31$). For the control group no significant increase in detection performance could be found (all $p > .10$, $d < .0.19$), see Table 5 for details. Training with XRT has an effect not only on the objects in the easy view but also on those in the difficult view. The screeners could make the association between the rotated object they detected during training and the canonical view of the object which is displayed in the object information in XRT.

In summary, a large and significant training effect was found for the group who trained with XRT for 3 and 6 months compared to a control group who used another CBT for the same time. A significant training effect has been observed for all four threat categories (guns, knives, IEDs and other), whereas the extent of the effect varied between categories. A large transfer of the acquired knowledge about the visual appearance of trained objects (set A) to untrained but similar looking objects (set B) was found for the XRT training group but not for the control group. This means that training with XRT helped screeners to detect other prohibited items which were not part of the XRT training. Substantial effects of viewpoint could be observed, i.e. unusual views of prohibited objects were much harder to detect than canonical views.

Table 5 Results of the t tests comparing the detection performance of the four categories for easy view (V1) and difficult view (V2) between the first (t1) and second (t2) measurement

	$t(96)$	$t(111)$	p Value	D
XRT training group				
Guns: V1t1–V1t2	−4.21		<0.01	0.53
IEDs: V1t1–V1t2	−12.25		<0.001	1.42
Knives: V1t1–V1t2	0.64		=0.53	0.07
Other: V1t1–V1t2	−8.95		<0.001	1.12
Guns: V2t1–V2t2	−6.03		<0.001	0.70
IEDs: V2t1–V2t2	−11.45		<0.001	1.43
Knives: V2t1–V2t2	−2.53		<0.05	0.31
Other: V2t1–V2t2	−6.17		<0.001	0.84
Control group				
Guns: V1t1–V1t2		−0.21	=0.84	0.02
IEDs: V1t1–V1t2		−0.76	=0.45	0.08
Knives: V1t1–V1t2		−0.66	=0.51	0.07
Other: V1t1–V1t2		−1.26	=0.21	0.13
Guns: V2t1–V2t2		−0.67	=0.50	0.09
IEDs: V2t1–V2t2		0.71	=0.48	0.07
Knives: V2t1–V2t2		1.65	=0.10	0.19
Other: V2t1–V2t2		0.64	=0.53	0.07

Experiment 2

The main aim of experiment 2 was to replicate the results of experiment 1 at another European airport. In addition, another conventional CBT was used for the control group. Thus it could be investigated whether conventional CBTs differ from each other regarding training effectiveness compared to XRT.

Method

Participants

A total of 163 airport security screeners of another mid-size European airport participated in experiment 2. All screeners conducted the X-ray CAT 1.0.0 three times with an interval of three months between the measurements. The adaptive CBT group (XRT group) consisted of 84 screeners who conducted weekly recurrent CBT using X-ray Tutor (XRT) CBS 2.0 Standard Edition between all three test measurements. The control group consisted of 79 screeners and they used another conventional CBT than the control group of experiment 1. As in experiment 1, according to the security organization and their appropriate authority, airport security screeners of both groups conducted about 20min CBT per week. Analysis of XRT training use showed that on average, each screener trained 20.92min (SD = 2.87) per week.

Materials and procedure

Materials and procedure in experiment 2 were the same as in experiment 1. Again, all screeners took the X-ray CAT at the beginning and after 3 and 6months of CBT. The only difference was the CBT for the control group, which was another one than in experiment 1. In order to avoid potential negative consequences, we decided not to mention the exact CBT product in this article for experiment 2, neither. However, it can be mentioned that also this CBT is widely used at many airports worldwide. As the conventional CBT used in experiment 1, this CBT has a much smaller threat image library than XRT, threat objects are not displayed in many different views, threat objects are not matched with different bags automatically on the fly, and there is no individually adaptive training algorithm.

Results and discussion

This section is structured the same way as in experiment 1. Figure 10 shows the detection performance d' for both groups and all three test measurements. As in experiment 1, individual d' scores were subjected to repeated measures ANOVA with the within-participant factor measurement (first, second and third) and the between-participant factor group (XRT training group and control group). Again, there were large main effects of measurement $\eta^2 = 0.50$, $F(2, 322) = 163.52$, $p < .001$, group, $\eta^2 = 0.26$, $F(1, 161) = 56.34$, $p < .001$, and a significant interaction of measurement and group $\eta^2 = 0.33$, $F(2, 322) = 78.40$, $p < 0.001$. The large interaction is consistent with Fig. 10 showing a much larger performance increase as a result of training for the XRT training group when compared to the control

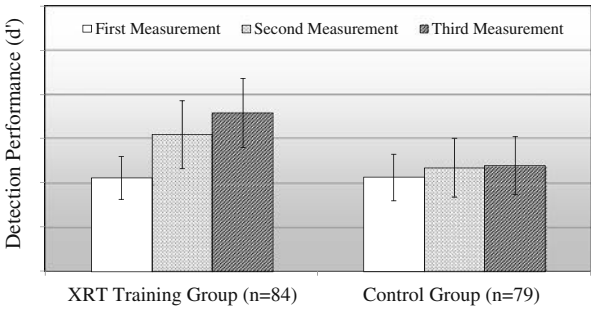

Fig. 10 Detection performance with standard deviations for the XRT training group vs the control group comparing first, second and third measurement

group. This was confirmed by independent samples *t* tests. There was no significant difference between both groups for the first measurement $t(161) = -0.22$, $p = 0.83$, $d = 0.03$, but a highly significant difference already in the second measurement $t(161) = 6.66$, $p < 0.001$, $d = 1.05$ after 3months of training. As in experiment 1, additional paired-samples *t* tests revealed significant differences for the XRT training group between all measurements. In contrast to experiment 1, there were also significant differences for the control group between the first and second measurement, although not between the second and third measurement (see Table 6). Thus, the conventional CBT used in experiment 2 did also result in increased detection performance although substantially less than XRT.

Figure 11 shows the detection performance of both screener groups broken up by prohibited item category and the three test measurements. Again, a clear effect of training on the detection performance can be seen for the XRT training group with the largest increase after the first 3months of training. However, also the control group shows a slight increase in detection performance at least for the second measurement. The analysis of variance (ANOVA) with threat category as additional within-participant factor showed significant main effects and significant interactions (for details see Table 7, a). The results are comparable to those in experiment 1. Most importantly, detection of guns was best initially, while detection of IEDs was much lower. After 6months of recurrent adaptive CBT, screeners of the XRT training group could detect IEDs even slightly better than guns. This nice replication of the results obtained in experiment 1 clearly shows that IED detection is not difficult per se but only a matter of the right training.. As mentioned above, all IEDs used in this study contained a detonator, wires, explosive, a triggering device and a power source. Thus our conclusions are only applicable to the detection of such multi-component IEDs. As shown in Table 8, *t* tests between the first and second measurement revealed significant training effects for the XRT training group for all

Table 6 Results of the *t* tests comparing the detection performance of first (*t*1), second (*t*2) and third (*t*3) measurement

	t(83)	*t*(78)	*p* Value	*D*
XRT training group (*t*1–*t*2)	−12.21		<0.001	1.57
XRT training group (*t*2–*t*3)	−7.07		<0.001	0.65
Control group (*t*1–*t*2)		−3.67	<0.001	0.36
Control group (*t*2–*t*3)		−0.91	=0.37	0.07

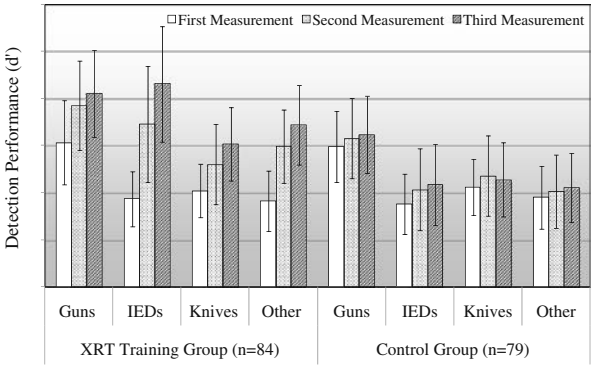

Fig. 11 Detection performance with standard deviations for the XRT training group vs the control group comparing first, second and third measurement for each threat category separately

threat categories with large effect sizes (all $d > 0.80$). In contrast to experiment 1, there were also significant effects for the control group, although with rather low effect sizes (all $d < 0.56$). Thus the conventional CBT used in experiment 2 also resulted in performance increases although much less than XRT.

By an ANOVA with measurement and set as within-participant factors and group as between-participants factor, we investigated if training effects can also be shown for threat objects which were not included in the training sessions. There were main effects and interactions for all factors showing similar results as in experiment 1 (see Table 7, b for details). As in experiment 1, a large transfer effect was found (see Fig. 12). Not only for the prohibited items of set A, which were included in the training library of XRT, but also for the untrained prohibited objects of set B, screeners of the XRT training group showed a large increase in detection performance after training. Paired-samples t tests between the first and second measurement showed training effects for both sets and also for both groups whereas again large effect sizes were found for the XRT training group and small effect sizes for the control group (trained group set A: $t(83) = -13.10$, $p < 0.001$, $d = 1.77$ and set B: $t(83) = -9.53$, $p < 0.001$, $d = 1.24$, control group set A: $t(78) = -2.32$, $p < 0.05$, $d = 0.24$ and set B: $t(78) = -3.00$, $p < 0.01$, $d = 0.32$). Pairwise t tests showed no significant difference in the difficulty of set A and Set B for both groups at the first measurement (XRT training group: $t(83) = 1.16$, $p = 0.25$, $d = 0.10$, control group: $t(78) = 1.93$, $p = 0.06$, $d = 0.19$).

Figure 13 includes also the threat category in the analysis. Paired samples t tests were calculated in order to investigate if the training effect between the first and second measurement was significant for each category in both sets for the XRT training group. Results revealed significant effects for all categories in each set ($p < 0.01$, $d = 0.51$ for knives in Set B, $p < 0.001$, $d > 0.74$ for all other categories). Thus, as in experiment 1, XRT resulted in large detection performance increases even for prohibited objects that are not part of the XRT image library (X-ray CAT image set B). For the control group the difference between the first and *third* measurement was calculated in order to maximize the chances for finding a significant training effect. The following t tests were significant: IEDs for both sets, knives only for set A, and other threat objects for both sets ($p < 0.05$, $d > 0.23$). All other values were not significant ($p > 0.06$, $d < 0.28$) and reveal no effect of training between the different measurements.

Table 7 Results of the ANOVAs in experiment 2

	Factor	*df*	*F*	η^2	*p* Value
a	Measurement (*M*)	2, 322	160.78	0.50	<0.001
	Threat category (*T*)	3, 483	234.85	0.59	<0.001
	Group (*G*)	1, 161	64.98	0.29	<0.001
	*M*x*G*	2, 322	78.54	0.33	<0.001
	*T*x*G*	3, 483	37.63	0.19	<0.001
	*M*x*T*	6, 966	26.24	0.14	<0.001
	*M*x*T*x*G*	6, 966	16.67	0.09	<0.001
b	Measurement (*M*)	2, 322	156.12	0.49	<0.001
	Set (*S*)	1, 161	58.45	0.27	<0.001
	Group (*G*)	1, 161	56.03	0.26	<0.001
	*M*x*G*	2,322	82.16	0.34	<0.001
	*M*x*S*	2, 322	8.88	0.05	<0.001
	*S*x*G*	1, 161	31.37	0.16	<0.001
	*M*x*S*x*G*	2, 322	15.52	0.09	<0.001
c	Measurement (*M*)	2, 322	162.28	0.50	<0.001
	Set (*S*)	1, 161	41.88	0.21	<0.001
	Threat category (T)	3, 483	231.83	0.59	<0.001
	Group (*G*)	1, 161	71.93	0.31	<0.001
	*M*x*G*	2, 322	84.18	0.34	<0.001
	*M*x*T*	6, 966	27.50	0.15	<0.001
	*M*x*S*	2, 322	11.42	0.07	<0.001
	*S*x*G*	1, 161	36.23	0.18	<0.001
	*S*x*T*	3, 483	33.59	0.17	<0.001
	*T*x*G*	3, 483	40.15	0.20	<0.001
	*M*x*T*x*G*	6, 966	16.87	0.10	<0.001
	*M*x*S*x*G*	2, 322	10.09	0.06	<0.001
	*M*x*S*x*T*	6, 966	1.48	0.01	=0.18
	*S*x*T*x*G*	3, 483	3.69	0.02	<0.05
	*M*x*S*x*T*x*G*	6, 966	2.64	0.02	<0.05
d	Measurement (*M*)	2, 322	152.62	0.49	<0.001
	View (*V*)	1, 161	1849.85	0.92	<0.001
	Threat category (*T*)	3, 483	216.74	0.57	<0.001
	Group (*G*)	1, 161	70.32	0.30	<0.001
	*M*x*G*	2, 322	80.05	0.33	<0.001
	*M*x*T*	6, 966	26.57	0.14	<0.001
	*M*x*V*	2, 322	2.99	0.02	=0.05
	*V*x*G*	1, 161	0.62	0.00	=0.43
	*V*x*T*	3, 483	288.98	0.64	<0.001
	*T*x*G*	3, 483	34.91	0.18	<0.001
	*M*x*T*x*G*	6, 966	14.95	0.09	<0.001
	*M*x*V*x*G*	2, 322	1.21	0.01	=0.30
	*M*x*V*x*T*	6, 966	2.82	0.02	<0.05
	*V*x*T*x*G*	3, 483	1.69	0.01	=0.17
	*M*x*V*x*T*x*G*	6, 966	1.89	0.01	=0.08

As in experiment 1, individual d' scores were subjected to an extended ANOVA with the within-participant factors measurement, X-ray CAT image set, threat category and the between-participants factor group. All main effects and interactions were significant except the interaction between measurement, set and threat category (see Table 7, c, for details). In contrast to experiment 1 the ANOVA revealed a main effect of set and significant interactions with set. However, as can be seen in Fig. 13 they were rather small, which implies large transfer effects. As in experiment 1 the

Table 8 Results of the *t* tests comparing the categories between first (*t*1), second (*t*2) and third (*t*3) measurement

XRT training group	*t*	*df*	*p* Value	*d*
Guns *t*1–*t*2	−6.01	83	<0.001	0.86
IEDs *t*1–*t*2	−12.84	83	<0.001	1.74
Knives *t*1–*t*2	−5.81	83	<0.001	0.80
Other *t*1–*t*2	−12.30	83	<0.001	1.64
Guns *t*1–*t*3	−8.19	83	<0.001	1.15
IEDs *t*1–*t*3	−20.22	83	<0.001	2.70
Knives *t*1–*t*3	−10.97	83	<0.001	1.48
Other *t*1–*t*3	−16.46	83	<0.001	2.18
Control group				
Guns *t*1–*t*2	−2.19	78	<0.05	0.23
IEDs *t*1–*t*2	−3.60	78	<0.01	0.42
Knives *t*1–*t*2	−2.73	78	<0.01	0.33
Other *t*1–*t*2	−1.46	78	<0.15	0.18
Guns *t*1–*t*3	−2.72	78	<0.01	0.34
IEDs *t*1–*t*3	−4.61	78	<0.001	0.56
Knives *t*1–*t*3	−2.05	78	<0.05	0.23
Other *t*1–*t*3	−2.59	78	<0.05	0.30

results clearly show a training effect for each category and in both sets. This is consistent with the results of the *t* tests explained above. The training effect that was found for the control group revealed itself also in the sets, that is, there was a transfer effect for the control group, too.

Last, the effect of viewpoint was investigated calculating a four-way ANOVA. Results show clear main effects of measurement, view, threat category and group. For details on interactions please refer to Table 7, d. Detection performance is clearly much higher for objects that are shown in the easy view (view 1) than for the objects that are shown from an unusual viewpoint (see Fig. 14). This effect is valid for all threat categories and for the XRT training group as well as for the control group. However, the viewpoint effect is not the same for different threat categories. The graphs in Fig. 14 suggest that the largest viewpoint effect can be observed for the detection of knives, the smallest one for IEDs.

As in experiment 1, pairwise *t* tests showed a significant increase in detection performance at the second measurement for both views for the XRT training group for all four threat categories ($p < 0.01$, $d > 0.49$. For the easy view, the control group showed a significant effect for IEDs only ($p < 0.05$, $d = 0.32$), all other *t* tests were not significant ($p > 0.07$, $d < 0.25$). For the difficult view all *t* test with one exception were

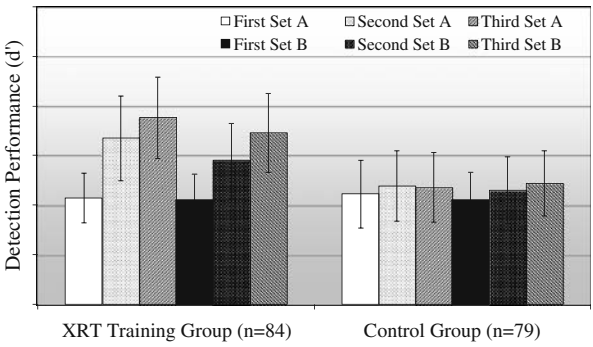

Fig. 12 Detection performance with standard deviations for the XRT training group vs the control group comparing first, second and third measurement for sets A and B separately

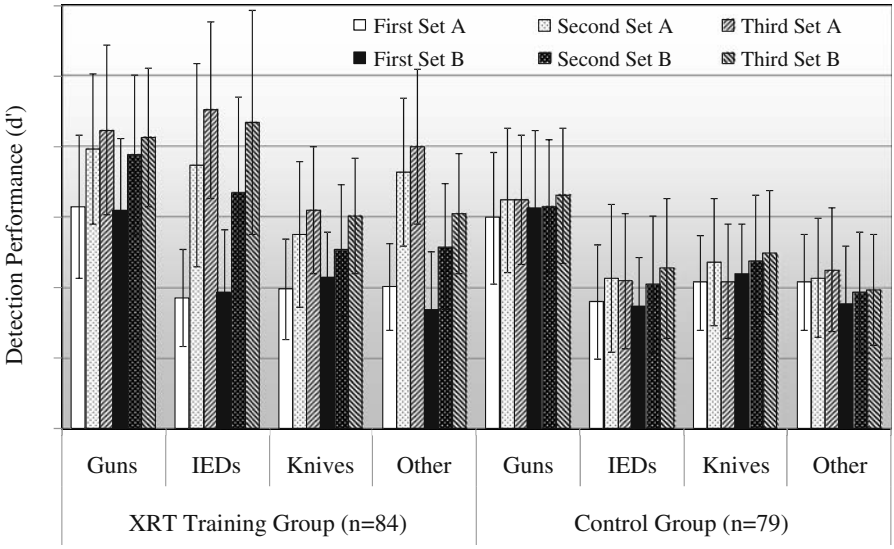

Fig. 13 Detection performance with standard deviations for the XRT training group vs the control group comparing first, second and third measurement for sets A and B and each threat category separately

significant for the control group ($p < 0.05$, $d > 0.26$). Only the training effect of knives in the rotated view was not significant $p = 0.07$, $d = 0.24$ (see Table 9 for details). But the results show that although some significant effects in the control group were observed, effect sizes were small compared to those of the XRT training group.

In summary, very similar results as in experiment 1 have been found in experiment 2. A large and significant training effect was observed for the group who trained with XRT

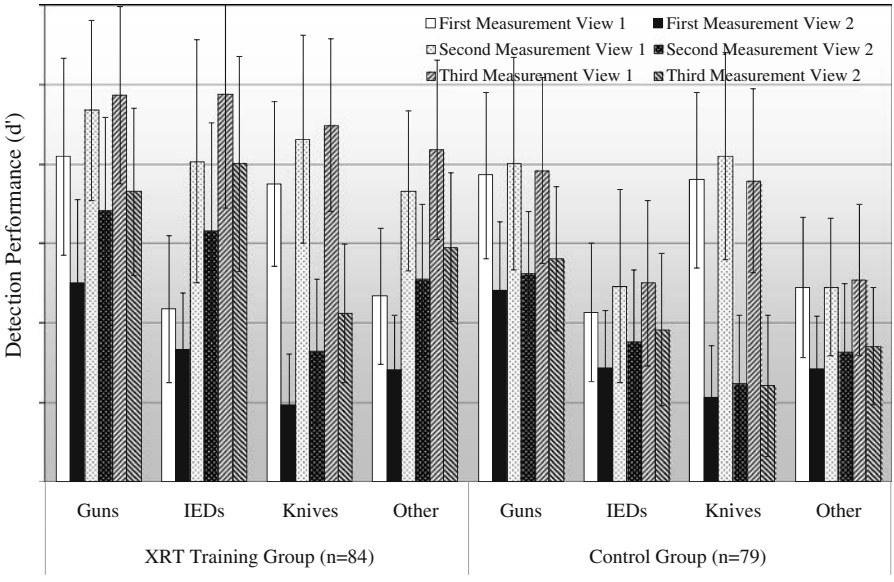

Fig. 14 Detection performance with standard deviations for the XRT training group vs the control group comparing first, second and third measurement for both views and each threat category separately

Table 9 Results of the t-tests comparing the detection performance of the four categories for easy view (*V*1) and difficult view (*V*2) between the first (*t*1) and second (*t*2) measurement

	t(83)	*t*(78)	*p* Value	*d*
XRT training group				
Guns: *V*1*t*1 − *V*1*t*2	−3.59		<0.01	0.49
IEDs: *V*1*t*1 − *V*1*t*2	−10.93		<0.001	1.51
Knives: *V*1*t*1 − *V*1*t*2	−4.35		<0.001	0.48
Other: *V*1*t*1 − *V*1*t*2	−9.79		<0.001	1.42
Guns: *V*2*t*1 − *V*2*t*2	−5.46		<0.001	0.82
IEDs: *V*2*t*1 − *V*2*t*2	−9.99		<0.001	1.45
Knives: *V*2*t*1 − *V*2*t*2	−5.79		<0.001	0.88
Other: *V*2*t*1 − *V*2*t*2	−10.33		<0.001	1.40
Control group				
Guns: *V*1*t*1 − *V*1*t*2		−1.07	=0.29	0.13
IEDs: *V*1*t*1 − *V*1*t*2		−2.64	<0.05	0.32
Knives: *V*1*t*1 − *V*1*t*2		−1.87	=0.07	0.25
Other: *V*1*t*1 − *V*1*t*2		−0.05	=0.96	0.01
Guns: *V*2*t*1 − *V*2*t*2		−2.35	<0.05	0.26
IEDs: *V*2*t*1 − *V*2*t*2		−3.24	<0.01	0.41
Knives: *V*2*t*1 − *V*2*t*2		−1.81	=0.07	0.24
Other: *V*2*t*1 − *V*2*t*2		−2.11	<0.05	0.28

compared to a control group who used a conventional CBT for the same time. A significant training effect has been observed for all four categories (guns, knives, IEDs and other) for the XRT training group, whereas the effect size varied between categories. Also a large transfer of the acquired knowledge about the visual appearance of trained objects (set A) to untrained but similar looking objects (set B) was found for the XRT training group. Additionally a viewpoint effect could be observed which shows that unusual views of forbidden objects are much harder to detect than canonical views. In contrast to experiment 1, the control group also showed increases of detection performance, which implies that the conventional CBT used in experiment 2 is more effective than the one used in experiment 1 (although still much less effective than XRT). Moreover, there was also a transfer effect for the control group.

General discussion

The first aim of this study was to investigate how well airport security screeners can detect guns, knives, IEDs and other prohibited items in X-ray images of passenger bags. Two experiments conducted at two European airports provided very similar results. A computer-based test (X-ray CAT) was conducted before and after 3 and 6months of weekly (about 20min per screener) CBT at each airport. The first measurement revealed that guns were detected best, followed by knives, other prohibited items and IEDs. In both experiments and airports, one group used an adaptive CBT (X-ray Tutor, XRT) with individually adaptive algorithms, a large library of prohibited items depicted in a variety of different views, and automatically created prohibited item to bag combinations (see Schwaninger 2004 for details). The other group used a conventional CBT system with no adaptive algorithms, a smaller image library, and fixed combinations of threat items in bags. While XRT was used in both experiments and airports, two different conventional CBT systems were used for the control groups of experiment 1 (airport 1) and experiment 2 (airport 2). At

both airports, XRT training group results revealed a training effect for all types of threat objects (guns, knives, IEDs, and other prohibited items). However, effect sizes differed remarkably for the four categories. While guns were detected best and IEDs were detected worst at the beginning, IED detection of the XRT training group was as good as or even slightly better than gun detection after several months of training. This shows that the detection of IEDs is not difficult per se, but rather depending on the training of screeners. Note that all IEDs used in this study contained a detonator, wires, explosive, a triggering device and a power source. Therefore, these conclusions are only applicable to the detection of such multi-component IEDs. However, a large training effect for IEDs can be expected because they are usually not encountered at airport security checkpoints and therefore not known to screeners without enhanced training in IED detection. The relatively large training effect for the category "other" which includes self defense gas spray, electric shock devices etc. might also be explained by less on the job exposure of these prohibited items. In a study with hold baggage screeners, large training effects for IEDs were also found, which is very consistent with results of this study (Schwaninger and Hofer 2004). In contrast to IEDs and other prohibited items, guns seem to be well known by screeners either because of their typical shape or the frequency by which they are encountered at the airport security screening checkpoint (e.g. toy guns). Therefore, detection performance before training is already high for guns and a large improvement is impossible. It is also noticeable that detection for knives showed the smallest training effect in both experiments. Although the detection was at the baseline measurement higher than for IEDs and other prohibited items, after six months of training screeners' performance was poorest for knives. On average, knives are smaller than IEDs and other threat items and show less diagnostic features. This might be a reason for the lower detection performance increase for this threat category.

While training with XRT resulted in large training effects, the tested conventional CBT systems were less effective. In experiment 1, there were no training effects at all, while only small training effects were observed for the conventional CBT system used in experiment 2. This could be due to one or a combination of the following reasons: First, the conventional CBT systems tested in this study do not feature individually adaptive training algorithms like XRT (see Schwaninger 2004 for details). Second, in contrast to XRT, the conventional CBT systems did not contain such a large image library with many prohibited items depicted from a variety of different viewpoints. Third, while in XRT prohibited items are blended into X-ray images of passenger bags on the fly using scientifically validated and individually adaptive algorithms based on image measurement as described in Schwaninger et al. (2007), the conventional CBT systems used in experiments 1 and 2 have only fixed combinations of prohibited items in bags. Finally, we had to rely on the statement of the appropriate authority and the security companies regarding the amount of training that was conducted by screeners of the control group and the XRT training group, which should have been on average 20min per week per screener. Analysis of XRT training data showed, that this was clearly fulfilled for screeners of the XRT training group at both airports.

Since the X-ray CAT is composed of two comparable (similar looking) sets (set A and set B) whereof only the threat objects of set A were included into the XRT

training system, transfer effects can be tested, i.e. whether training with certain prohibited items helps increasing detection of other prohibited items that are not contained in the training. Overall, the comparison of the two sets A and B at the baseline measurement (before training) shows no significant difference. However, in experiment 1 there was a slight difference for the control group between the two sets indicating that the two sets are not exactly equal in terms of image difficulty for this sample. But this possible objection to the transfer effect can be disapproved with two arguments: first, the effect size was only small according to the conventions by Cohen (1988) and second, only one of the two control groups showed a significant difference. Therefore, the transfer effect in the results of the XRT training group can be attributed to the training of set A only. The small training effect for the control group in experiment 2 is also reflected in the detection increase of both sets after training. Although the conventional CBT system of this control group did not contain any objects from the test, the training with this training system apparently also led to a transfer of the knowledge to the objects in the test. In another study it would be interesting to compare the objects that are comprised in the two training systems used by the control groups regarding their similarity to the test objects. Contrary to our results, Smith et al. (2005) found a large decrease in screeners' detection performance when specific trained objects were replaced with new images belonging to the same categories (p. 458; see also Smith et al. 2005, p. 1181). According to these authors, improvement in screening performance is attributable only to specific-token familiarity that developed for the original images and not to a category generalization. They state constraints on categorization and the use of category-general information when humans face visual complexity and have to identify targets within it. Our results can be interpreted in support of generalization of visual learning in X-ray image interpretation. However, it might be possible that the objects of the untrained set in our study are so similar to the trained objects that a specific-token familiarity led to the detection performance increase and not a true generalization effect. The lacking transfer effect in knives would along these lines mean that the objects in sets A and B are not similar enough in shape to generate a specific-token familiarity. Therefore only the learnt objects could generate a training effect but not the unlearnt ones. For Schwaninger and Hofer (2004) findings of a large increase in detection performance of IEDs after recurrent CBT with other members of the category than those included in the test, it would mean, that those objects were very similar in order to create a specific-token familiarity and therefore a training effect.

In both experiments a large viewpoint effect was also revealed. This is consistent with view-based theories of object recognition (for reviews see for example Tarr and Bülthoff 1995, 1998; Graf et al. 2002; Hayward 2003). After training, easy and difficult views were recognized much better. Interestingly, there was no significant interaction between measurement and viewpoint, i.e. although training resulted in improved performance for difficult views, the viewpoint effect (impairment for unusual vs canonical views) remained stable even after 6months of training. However, it must be pointed out that the XRT training algorithm only provides the screeners with unusual views of objects once a screener can detect a prohibited item well when depicted from easy perspective. That is, when screeners start to train with XRT all threat objects are shown in easy views. Only if these objects are detected reliably, the difficulty level is

increased for a certain threat item by showing it in more difficult views (Schwaninger 2004). Thus, it is unclear whether a significant interaction between viewpoint and measurement would have been observed if the training duration would have been increased (e.g. to 1year). The conclusion stands to reason that recognition of forbidden objects in X-ray images is dependent on exposure which has very important implications for an adaptive training system. It has been assumed that different views of each object become associated with one another during object rotation, either through active learning or through passive experiencing of the successive appearance of nearby views (Földiák 1991; Stryker 1991). Hence, it is important that during training screeners are getting feedback which forbidden object has been detected or missed. This feedback shows the photograph and also the X-ray image of that forbidden object always in the canonical view whereas the forbidden object merged into a bag is presented in different viewpoints. This leads to an association between an unusual view of an object and the canonical view which results in a sequential pairing of these views with each other (Wang et al. 2005). This association, which forms during learning, is thought to underlie object recognition ability across changes in viewing angle (Palmeri and Gauthier 2004).

For our future studies, it could also be interesting to increase the interval between the end of training and the testing of training transfer, as corresponding literature usually tests transfer of training after a considerable period of time in order to measure the stability of the transfer (e.g., Saks and Belcourt 2006). In any case, our findings show that the knowledge about the visual appearance of forbidden objects, which airport security screeners acquire during recurrent CBT, can be transferred to similar looking, but not previously seen objects and also the effect that rotated views are much harder to detect can be decrease with training. To make sure that objects are well detected it is important that a large and representative image library of prohibited objects is used and that these objects are learned from different viewpoints. Additionally the library should be updated constantly to adapt to new threats. Overall, this study has shown that adaptive CBT can be a powerful tool to increase screeners' X-ray image interpretation competency in an efficient and effective way.

Acknowledgment This research was financially supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403).

References

Bülthoff HH, Edelman SY, Tarr MJ (1995) How are three-dimensional objects represented in the brain? Cereb Cortex 3:247–260

Biederman I (1987) Recognition-by-components: a theory of humanimage understanding. Psychol Rev 94:115–147

Cohen J (1988) Statistical power analysis for the behavioural sciences. Erlbaum, Hillsdale, New York

Földiák P (1991) Learning invariance from transformation sequences. Neural Comput 3:194–200

Graf M, Schwaninger A, Wallraven C, Bülthoff HH (2002) Psychophysical results from experiments on recognition and categorisation. Information Society Technologies (IST) Programme, Cognitive Vision Systems–CogVis (IST-2000-29375)

Green DM, Swets JA (1966) Signal detection theory and psychophysics. Wiley, New York

Hardmeier D, Hofer F, Schwaninger A (2006) The role of recurrent CBT for increasing aviation security screeners' visual knowledge and abilities needed in X-ray screening. The 4th International Aviation Security Technology Symposium, Washington, DC, USA, November 27–December 1, 2006

Hayward WG (2003) After the viewpoint debate: where next in object recognition? Trends Cogn Sci 7:10425–427

Hummel JE, Biederman I (1992) Dynamic binding in a neural network for shape recognition. Psychol Rev 99(3):480–517

Koller S, Schwaninger A (2006) Assessing X-ray image interpretation competency of airport security screeners. Proceedings of the 2nd International Conference on Research in Air Transportation, ICRAT 2006 Belgrade, Serbia and Montenegro, June 24–28, 2006:399–402

Kosslyn SM (1994) Image and brain. The resolution of the imagery debate. MIT Press Cambridge, MA

Lawson R (1999) Achieving visual object constancy across plane rotation and depth rotation. Acta Psychol 102:221–245

Logothetis NK, Sheinberg DL (1996) Visual object recognition. Annu Rev Neurosci 19:577–621

Lowe D (1985) Perceptual organization and visual recognition. Kluwer Academic, Boston

Lowe DG (1987) Three-dimensional object recognition from single two-dimensional images. Artif Intell 31:355–395

Marr D, Nishihara HK (1978) Representation and recognition of the spatial organization of three-dimensional shapes. Proc R Soc Lond B Biol Sci 200:269–294

McCarley JS, Kramer AF, Wickens CD, Vidoni ED, Boot WR (2004) Visual skills in airport screening. Psychol Sci 15(5):302–306

Palmer SE, Rosch E, Chase P (1981) Canonical perspective and the perception of objects. In: Long I, Baddeley A (eds) Attention and performance IX. Erlbaum, Hillsdale, NJ

Palmeri TJ, Gauthier I (2004) Visual object understanding. Nat Rev Neurosci 5:291–303

Peissi J, Tarr MJ (2007) Visual object recognition: do we know more now than we did 20 years ago? Annu Rev Psychol 58:75–96

Poggio T, Edelman S (1990) A network that learns to recognize three-dimensional objects. Nature 343 (6255):263–266

Saks AM, Belcourt M (2006) An investigation of training activities and transfer of training in organizations. Hum Resour Manag 45(4):629–648

Schwaninger A (2003) Training of airport security screeners. Airport 05/2003:11–13

Schwaninger A (2004) Computer-based training: a powerful tool to the enhancement of human factors. Aviation Security International FEB/2004:31–36

Schwaninger A (2005a) X-ray imagery: enhancing the value of the pixels. Aviation Security International 2005:16–21

Schwaninger A (2005b) Object recognition and signal detection. In: Kersten B (ed) Praxisfelder der Wahrnehmungspsychologie. Huber, Bern

Schwaninger A, Hardmeier D, Hofer F (2005) Aviation security screeners visual abilities and visual knowledge measurement. IEEE Aerosp Electron Syst 20(6):29–35

Schwaninger A, Hofer F (2004) Evaluation of CBT for increasing threat detection performance in x-ray screening. In: Morgan K, Spector MJ (eds) The Internet Society 2004, Advances in Learning, Commerce and Security. WIT Press, Wessex

Schwaninger A, Hofer F, Wetter OE (2007) Adaptive computer-based training increases on the job performance of x-ray screeners. Proceedings of the 41st Carnahan Conference on Security Technology, Ottawa, October 8–11, 2007. Exp Psychol Gen 134(4):443–460

Schwaninger A, Michel S, Bolfing A (2007) A statistical approach for image difficulty estimation in x-ray screening using image measurements. Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization. ACM Press, New York, USA123–130

Smith JD, Redford JS, Washburn DA, Taglialatela LA (2005) Specific-token effects in screening tasks: possible implications for aviation security. J Exper Psychol Learn Mem Cogn 31(6):1171–1185

Stryker MP (1991) Temporal associations. Nature 354:108–109

Tarr MJ (1995) Rotating objects to recognize them: a case study on the role of viewpoint dependency in the recognition of three-dimensional objects. Psychon Bull Rev 2:55–82

Tarr MJ, Bülthoff HH (1995) Is human object recognition better described by geon-structural-descriptions or by multipleviews? J Exp Psychol Hum Percept Perform 21(6):1494–1505

Tarr MJ, Bülthoff HH (1998) Image-based object recognition in man, monkey and machine. In: Tarr MJ, Bülthoff HH (eds) Object recognition in man, monkey, and machine (1–20). MIT Press, Cambridge, MA

Tarr MJ, Pinker S (1989) Mental rotation and orientation-dependence in shape-recognition. Cogn Psychol 21(2):233–282

Ullman S (1998) Three-dimensional object recognition based on the combination of views. In: Tarr M, Bülthoff H (eds) Object recognition in man, monkey and machine. MIT Press, England, pp21–44

Verfaillie K (1992) Variant points of view on viewpoint invariance. Can J Psychol 46:215–235

Wang G, Obama S, Yamashita W, Sugihara T, Tanaka K (2005) Prior experience of rotation is not required for recognizing objects seen from different angles. Nat Neurosci 8(12):1768–1775